Phonetica
Editor: K. Kohler, Kiel
Publishers: S. Karger, Basel
Separatum (Printed in Switzerland)

Phonetica *38:* 9–20 (1981)

A Temporal Model of Speech Production

FREDERICKA BELL-BERTI and KATHERINE S. HARRIS

St. John's University, Jamaica, N.Y.; Haskins Laboratories, New Haven, Conn.;
The Graduate School, The City University of New York, New York, N.Y., USA

Abstract. Existing models of speech production and coarticulation have failed to account for observations of real speech because they have considered timing to be a by-product of articulatory events instead of an integral organizing parameter of the speech motor plan. The model offered here considers time and timing relationships to be intrinsic to speech motor organization and the units of speech to be inherently dynamic gestures rather than static vocal tract configurations or invariant commands to the articulators.

Attempts to model speech production processes, particularly coarticulation, have failed to account for a number of observations of natural speech, perhaps because they have failed to consider timing as a basic parameter of the speech motor plan. Although time specification is only marginally part of conventional articulatory phonetics [LISKER, 1974], we believe that the temporal relationships among the articulatory components of a particular segment are integral to speech organization. Before putting forward a detailed description of this view, however, we will summarize the existing classes of models and provide brief accounts of their shortcomings in predicting temporal events.

Almost all of the existing speech production and coarticulation models are organized in terms of conventional linguistic/phonetic units – that is, it is assumed that the rules for coarticulation can be specified by writing some representation of the units on a page, and indicating coarticulatory boundaries, although the types of units to be written differ in their type; in some models (top-down models), the larger units (e.g. syllables, stressed vowels) dominate the smaller, while in others (bottom-up models), smaller units determine characteristics of the larger chunks by feature-spreading or feature-exchange between

adjacent segments. In both these types of model, the values assumed by nonessential articulatory features are mutable; indeed, it is this property that allows the speaker to make successive segments compatible within each new phonetic context, and it is the resulting changes that are traditionally identified as coarticulation. In all these views, the internal temporal characteristics of the segment are unrepresented. In all these models, while segment characteristics are presumed to have temporal consequences, temporal specification is not part of the segment description.

The former type of model may be exemplified by that of KOZHEVNIKOV and CHISTOVICH [1965], which postulates that, within the *syntagma*, speech is organized into *articulatory syllables* having a C_nV structure. For example, the phonetic feature 'lip rounding' is described as spreading to the first member of a string of nonlabial consonants preceding a vowel, regardless of the morpheme boundary conventionally defining the syllable and regardless of the number of segments in the string [DANILOFF and MOLL, 1968; BENGUEREL and COWAN, 1974]. Although the work of BENGUEREL AND COWAN on lip rounding has usually been cited as support for such an articulatory syllable, these authors themselves note that anticipatory coarticulation can begin during the preceding vowel. These models also fail to predict the anticipation of nasality across such vowel sequences as 'free Ontario' [MOLL and DANILOFF, 1971] because of the anticipation over adjacent vowels. Thus, the general conclusion would appear to be that while anticipatory coarticulation spreads over several segments, boundaries cannot be easily specified by phonetically motivated syllabic chunks.

Another type of top-down model with an inadequate representation of temporal parameters was originally proposed by ÖHMAN [1966] and adopted in part by PERKELL [1969] and FOWLER [1977]. In this model, coarticulation stretches from stressed vowel to stressed vowel. The reason advanced for the coarticulatory spread depends, however, on the complementary segmental features of vowels and consonants – that since vowel production requires a global shaping of the vocal tract while consonant effects are more localized, coproduction is possible between vowel and consonant, but not between vowel and vowel.

One hypothesis of such models is that speech is organized around stressed vowels and that consonant and unstressed vowel gestures are superimposed upon (coproduced with) one or another stressed vowel. This model predicts that, in a VC_nV utterance, the time from the

beginning of the first vowel to the end of the second vowel will be relatively constant across large changes in the duration of the medial consonant cluster, because the duration of the consonant string is always coarticulated with a preceding or following vowel. This is, in fact, just the reverse of what actually happens: As the consonant string increases in duration, the time between the beginning of the first vowel and the end of the second vowel increases. Indeed, examining measurements of acoustic segment durations for 3 speakers producing between 10 and 15 repetitions each of from 12 to 18 different utterance types, we found correlations of from $r = 0.85$ to $r = 0.91$ between the durations of the medial consonant string and the entire VCV[1]. That is, although there is some shortening of the vowels as the consonant string duration increases (the correlations are not perfect, indicating that the relationship is not linear), the increases in overall duration imply that the consonant strings are not simply occupying more of the 'acoustic time' of the stressed vowels, while the relative timing of the latter remains unchanged. Instead, the timing of the stressed vowel gestures changes as the duration of the medial consonant string changes.

Given the problems in specifying temporal characteristics of coarticulation, in such 'top-down' models of articulatory organization, we may consider the alternative 'bottom-up' of feature-based models such as that of HENKE [1966]. He assumes that the input to the articulatory system is a string of phonemes that are specified as sets of invariant articulatory 'goals' or 'features'. This model postulates a 'look-ahead' procedure that allows the goals of phonemes occurring later in the string to influence the current and intervening vocal tract configurations, so long as these anticipated goals are not in conflict with any more immediate goals. Thus, one would expect to see anticipation of compatible features as early in preceding segment strings as possible, e.g., nasality should be anticipated in a string of vowels before a nasal consonant [MOLL and DANILOFF, 1971]. The feature specification has no temporal extent, but merely inheres in segments.

However, while feature-based models have been said to account for anticipatory nasal coarticulation as in the 'free Ontario' example cited

[1] These measurements were made on the data set reported in BELL-BERTI and HARRIS [in preparation] of the string [əlVC$_n$Vlə]. A possible quibble with the results is that the increase of duration occurs because of some characteristic of coarticulation of the string with the initial and final CVs. This possibility seems rather far-fetched, however, and there is no experimental support for it.

beginning of the first vowel to the end of the second vowel will be relatively constant across large changes in the duration of the medial consonant cluster, because the duration of the consonant string is always coarticulated with a preceding or following vowel. This is, in fact, just the reverse of what actually happens: As the consonant string increases in duration, the time between the beginning of the first vowel and the end of the second vowel increases. Indeed, examining measurements of acoustic segment durations for 3 speakers producing between 10 and 15 repetitions each of from 12 to 18 different utterance types, we found correlations of from $r = 0.85$ to $r = 0.91$ between the durations of the medial consonant string and the entire VCV[1]. That is, although there is some shortening of the vowels as the consonant string duration increases (the correlations are not perfect, indicating that the relationship is not linear), the increases in overall duration imply that the consonant strings are not simply occupying more of the 'acoustic time' of the stressed vowels, while the relative timing of the latter remains unchanged. Instead, the timing of the stressed vowel gestures changes as the duration of the medial consonant string changes.

Given the problems in specifying temporal characteristics of coarticulation, in such 'top-down' models of articulatory organization, we may consider the alternative 'bottom-up' of feature-based models such as that of HENKE [1966]. He assumes that the input to the articulatory system is a string of phonemes that are specified as sets of invariant articulatory 'goals' or 'features'. This model postulates a 'look-ahead' procedure that allows the goals of phonemes occurring later in the string to influence the current and intervening vocal tract configurations, so long as these anticipated goals are not in conflict with any more immediate goals. Thus, one would expect to see anticipation of compatible features as early in preceding segment strings as possible, e.g., nasality should be anticipated in a string of vowels before a nasal consonant [MOLL and DANILOFF, 1971]. The feature specification has no temporal extent, but merely inheres in segments.

However, while feature-based models have been said to account for anticipatory nasal coarticulation as in the 'free Ontario' example cited

[1] These measurements were made on the data set reported in BELL-BERTI and HARRIS [in preparation] of the string [əlVC$_n$Vlə]. A possible quibble with the results is that the increase of duration occurs because of some characteristic of coarticulation of the string with the initial and final CVs. This possibility seems rather far-fetched, however, and there is no experimental support for it.

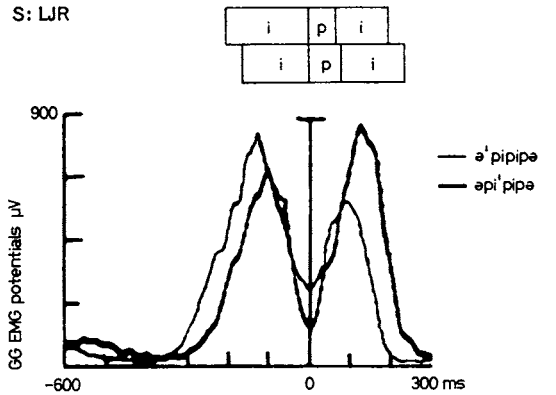

Fig. 1. Ensemble-averaged EMG potentials from the genioglossus (GG) muscle. Zero on the abscissa represents the point in the acoustic signal chosen for aligning tokens for averaging, in these utterances closure for /p/. Note the 'trough' in each trace.

above [MOLL and DANILOFF, 1971; for evidence to the contrary, see USHIJIMA and HIROSE, 1974], they fail to predict other coarticulatory phenomena, such as the suppression of tongue-fronting activity for a vowel during the production of an intervening consonant that contains no conflicting feature specification. For example, one specific prediction made by such a model is that the tongue gestures for the vowels in the sequence [ipi] should be continuous, since there is nothing in the unrounded vowel gesture which conflicts with the production of the labial consonant. In fact, the activity is not continuous [BELL-BERTI and HARRIS, 1974]; instead, the cessation of genioglossus muscle EMG activity between the vowels is striking (fig. 1). Furthermore, the extent of this suppression of EMG activity is highly correlated with duration of the closure for the medial consonant[2]. While it might be possible to claim that the consonant has some characteristic not actualized in its conventional feature description that causes the discontinuity, there is something unattractive about such an *ad hoc* solution.

In short, none of these models of coarticulation fits all the experimental facts. FOWLER [1980] has argued that their failure is inevitable, given that the models all belong to the class of extrinsic timing theories,

[2] Parallel fluorographic observations also reveal this discontinuity in vowel-to-vowel tongue gestures [BAER, personal commun.]. This same cessation of vowel EMG activity in a 'secondary' articulator, the lips, has been reported by BELL-BERTI and HARRIS [in preparation] and by GAY [1978].

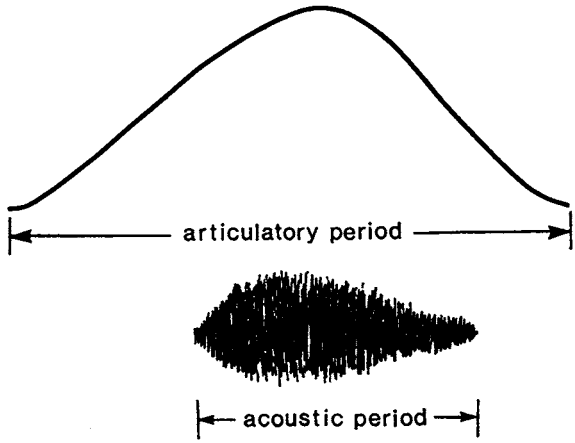

Fig. 2. The 'articulatory period' begins before and ends after the 'acoustic period'.

in which the units, whatever they may be, show an articulatory plan as an ordinal sequence without a true time axis. Here, a specific temporally-based model is proposed with some specific experimental consequences.

The model offered here considers time and timing relationships to be intrinsic to speech motor organization and the units of speech to be dynamic gestures rather than static vocal tract configurations. This model is derived from data collected during the last several years, in a series of experiments designed to provide parallel data from several articulators, using a variety of measurement techniques. These data are reported in detail elsewhere; we offer here only examples that illustrate various characteristics of this temporally based model.

We will focus first on describing the within-segment and then on the between-segment timing relationships proposed by the model. We believe that these relationships can account for the coarticulatory data in the literature as well as for our own data. The model proposes three rules for the timing of articulatory activity underlying segment representation. First, the articulatory period of a segment is longer than its acoustic period. That is, movements toward and away from the conventionally described articulatory goals for a segment begin before, and end after, the acoustic period of the segment. Thus, the movements toward an articulatory goal are considered to be as much a part of the specification of the segment as is the goal itself (fig. 2).

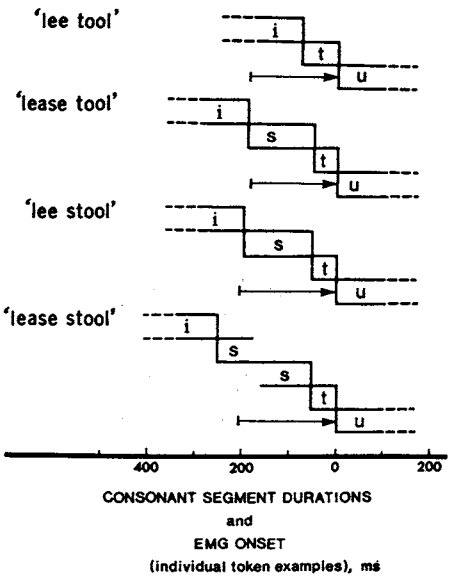

Fig. 3. The boxes represent the acoustic durations of the phones indicated: for /s/, they represent the duration of friction noise; for /t/, they represent the duration of the closure interval. The arrows indicate the times, before [u], at which orbicularis oris activity for the rounding gesture was first observed in each of these tokens.

The second rule is that for a given articulator the period of anticipation is temporally independent of preceding phone string length, if there is no articulatory conflict. Thus, in the often cited case of anticipatory lip rounding, rounding begins at a constant time before the vowel with which it is associated in a segmental string in which it is preceded by varying numbers of nonlabial consonants. The nonlabial consonants do not acquire a rounding feature; rather, there is cooccurrence of vowel-associated lip rounding with tongue activity associated with the consonant string. Since consonants tend to shorten in clusters, feature-spreading and simple cooccurrence cannot be distinguished without extensive systematic comparison of physiologic and acoustic measures. For example, we have reported that EMG activity from the orbicularis oris muscle, active in rounding and protruding the lips for the vowel [u], begins a constant time before the beginning of the acoustic period of the vowel [BELL-BERTI and HARRIS, 1979, in preparation]. Figure 3 offers several examples from a much larger data set. Clearly, in spite of differences in the durations of the consonant strings, the

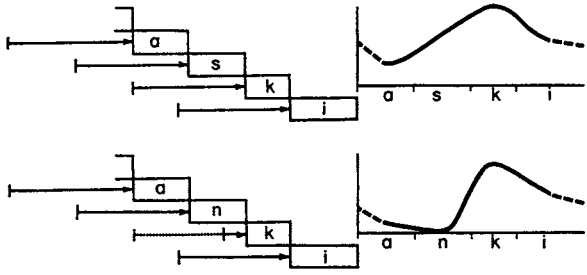

Fig. 4. Acoustic period durations of the segments in [aski] and [anki] are represented by boxes at the left. The beginning of each velar articulatory period is shown by arrows beginning to the left of each segment's acoustic period. The broken line to the left of [k] indicates suppression of velar elevation until after adequate port opening has been achieved for [n]. The resulting patterns of velar elevation are shown at the right, with the acoustic periods of the segments indicated along the abscissa.

number of consonants in each string, and the location of syllable boundaries within the strings, for this speaker EMG activity begins at about 200 ms before the vowel. We are not claiming that this relationship is constant across speakers or across speaking rate, although for one of our speakers the timing relationship between the beginning of tongue-fronting EMG activity and the beginning of the acoustic period of the vowel is very stable across changes in both speaking rate and lexical stress [BELL-BERTI, 1979.]

However, when there is conflict between the specification of proximate segments, the model presumes that realization of the later specification is altered. This possibility is exemplified in figure 4, where velar elevation for /k/ is delayed in the sequence /anki/ until an adequately low velar position has been achieved for the preceding /n/.

The central problem in this formulation is the definition of 'conflict'. Conventional feature specifications of phones are often inadequate as descriptions of articulatory realizations. For example, it has long been known that high and low oral vowels are realized with different velar heights [FRITZELL, 1969]. Such differences have coarticulatory consequences in velar height [BELL-BERTI et al., 1979]. While the spatial variations for velum height, apart from the feature specification 'oral' or 'nasal', have been fairly extensively studied, this is less true for features associated with other articulators. For example, no specification of lip position is included in feature descriptions of alveolar consonants, although some lip positions may be precluded for some consonants.

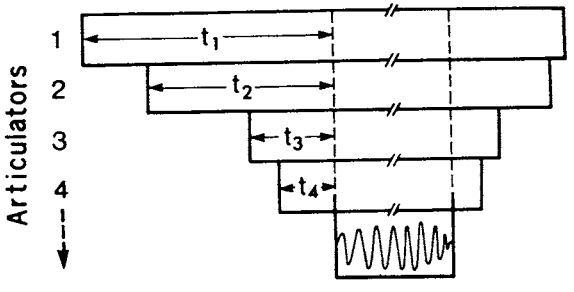

Fig. 5. This figure summarizes the three within-segment timing relationship properties of the model. First, the articulatory gestures begin before the segment's acoustic period begins. Second, each articulator's 'onset time' (the time between the beginning of the articulatory period and the beginning of the acoustic period) is constant. Third, 'onset time' is articulator-specific.

The third, and final, within-segment temporal rule specified by the model is that the articulatory period may begin at different times for different articulators (fig. 5). Although we do not know whether such nonsimultaneity should be attributed to such low-level properties of the system as articulatory biomechanics or to characteristics of segment motor plans, the durations of the parts of the articulatory period preceding and following the acoustic period must be empirically determined for each articulator. Similarly, the identity of the articulators themselves must be empirically determined. For example, it is possible that the organizational structure of the speech motor plan may include such traditionally described units as the velum and tongue tip and lips. Alternatively, the structural units may be functional groups of muscles, such as those acting to lower the jaw or move the tongue horizontally.

For example, BAER and ALPHONSO [1980] describe differences in the onset time of horizontal tongue movement between fronted and retracted vowels, with retracting movements for back vowels beginning earlier than the corresponding fronting movements for front vowels. This suggests that the different muscle groups acting to front and retract the tongue body may be the relevant structural units, and that they may have different onset times. In addition, the notion that the structural units of the motor program are functional groups of muscles allows for those cases in which one muscle may have two quite different functions when acting in conjunction with different muscle groups [see COLLIER et al., in preparation, for a discussion of such apparent dif-

ferences in mylohyoid function]. We might also add that the notion of variable functional groupings of muscles is compatible with some powerful theoretical models of speech motor control [FOWLER, et al. 1980].

A major consequence of articulator-specific and constant onset times is that the timing relationships among the articulators active for a given segment are fixed, although not simultaneous. This expectation is supported not only by our own data on lip rounding [BELL-BERTI and HARRIS, 1979, in preparation] and velar coarticulation [BELL-BERTI, in press], but also by data of KENT et al. [1974] on velar control timing.

These rules generate a timing rule in the model for segment strings. Quite simply, it is postulated that there is a constant overlap between the end of the articulatory period of one segment and the beginning of the articulatory period of the next. We have recently examined the timing relationships between the beginning of genioglossus muscle activity for /i/ and the end of orbicularis oris muscle activity for /p/, assuming that these measures reflect the beginning and end of the articulatory periods of the segments, respectively [RAPHAEL, 1975; TULLER and HARRIS, 1980; BELL-BERTI, 1979]. In one experiment in this series, the utterances all contained the sequence /pi/, although in different positions within a longer phrase, and with different lexical stress and syllable boundaries and at fast and slow speaking rates. The range of values of these intervals between the end of orbicularis oris EMG activity for /p/ and the beginning of genioglossus EMG activity for /i/ was narrow (fig. 6), and the distribution of these intervals about their mean is normal, and the entire range is within the time-constant used in integrating the EMG data (35 ms). (The extent of this overlap is expected to decrease at such points as temporally marked junctural boundaries).

We are not alone in noting those problems in describing speech motor organization which arise from a neglect of the temporal domain. For example, KENT and MINIFIE [1977], reflecting on the observations of KENT et al. [1974], suggest that a hierarchical model may be necessary to account for speech production, with (evidently) time-free specification of upstream stages of the model and time introduced in articulatory realization stages. While their plan is attractive, we believe that it is essential to avoid models of a code-decode type; that is, articulation should not be viewed as if it were the realization of cortically issued motor commands for static phonological units or their com-

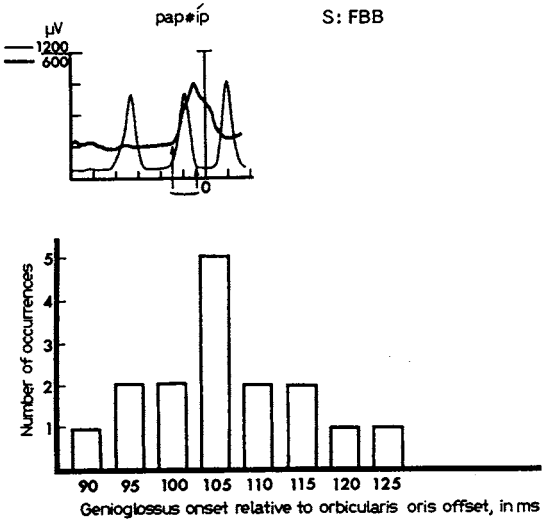

Fig. 6. Top: Ensemble-averaged EMG potentials from the orbicularis oris muscle (——) and the genioglossus muscle (——) for the disyllable [pap'ip]. The left arrow indicates the time at which genioglossus EMG activity began; the right arrow indicates the time at which orbicularis oris EMG activity ended. The interval between these points was measured for the ensemble averages of 20 repetitions of 16 utterances, 8 spoken at normal rate, 8 spoken rapidly. The utterances had the form /pVpVp/; the vowels were /i/ and /a/ or /a/ and /i/; stress was on the first or second vowel; the syllable boundary fell before or after the medial /p/. *Bottom:* Normal distribution of these intervals, i.e. the time from the beginning of genioglossus EMG activity to the end of orbicularis oris EMG activity.

ponent features, which suffer a recoding process in the periphery. It seems to us that the notion that speech production and perception must be linked by an encoding and a decoding process is theoretically faulty [FOWLER et al., 1980]. Rather, we believe that phonological units are inherently dynamic and retain this essential property when they are produced. Thus, the common failure to uncover static units, either as invariant commands to the articulators [MACNEILAGE and DECLERK, 1969] or as static articulator positions [LINDBLOM, 1963], does not indicate that speech production obscures the static properties of segments, but rather that the articulatory stream has been described in the wrong way. We hope that systematic observations, of the sort described in this paper, will allow us to formulate units that are more natural to a motor process occurring in time than those we have presently borrowed from phonological theory.

Acknowledgment

This work was supported by NINCDS grants NS–13617 and NS–05332 and by BRS grant RR–05596. We are grateful to BETTY TULLER for her suggestions on this manuscript.

References

BAER, T.; ALPHONSO, P.J.: Simultaneous cineradiographic-electromyographic-acoustic analysis of vowel production (Abstract). J. acoust. Soc. Am. *67:* 593 (1980).

BELL-BERTI, F.: Syllables and segments: a study of inter-articulator timing (Abstract). Asha *21:* 697 (1979).

BELL-BERTI, F.: Velopharyngeal function: a spatial-temporal model; in LASS Speech and language: advances in basic research and practice, vol. IV (Academic Press, New York, in press).

BELL-BERTI, F.; BAER, T.; HARRIS, K.S.; NIIMI, S.: Coarticulatory effects of vowel quality on velar elevation. Phonetica *36:* 187–193 (1979).

BELL-BERTI, F.; HARRIS, K.S.: More on the motor organization of speech gestures. Haskins Lab. Status Rep. Speech Res., SR 37/38, pp. 73–77 (Haskins Laboratories, New Haven 1974).

BELL-BERTI, F.; HARRIS, K.S.: Anticipatory coarticulation: some implications from a study of lip rounding. J. acoust. Soc. Am. *65:* 1268–1270 (1979).

BELL-BERTI, F.; HARRIS, K.S.: Temporal patterns of coarticulation: lip rounding (in preparation).

BENGUEREL, A.-P.; COWAN, H.A.: Coarticulation of upper lip protrusion in French. Phonetica *30:* 41–55 (1974).

COLLIER, R.; BELL-BERTI, F.; RAPHAEL, L.J.: The dynamics of Dutch diphthongs (in preparation).

DANILOFF, R.G.; MOLL, K.L.: Coarticulation of lip-rounding. J. Speech Hear. Res. *11:* 707–721 (1968).

FOWLER, C.: Timing contol in speech production (Indiana University Linguistics Club, Bloomington, Ind. 1977).

FOWLER, C.: Coarticulation and theories of extrinsic timing control. J. Phonet. *8:* 113–133 (1980).

FOWLER, C.A.; RUBIN, P.; REMEZ, R.E.; TURVEY, M.T.: Implications for speech production of a general theory of action; in BUTTERWORTH Language production (Academic Press, New York 1980).

FRITZELL, B.: The velopharyngeal muscles in speech: an electromyographic and cineradiographic study. Acta oto-lar., suppl. 250 (1969).

GAY, T.J.: Articulatory units: segments or syllables? in BELL, HOOPER Syllables and segments (North-Holland, Amsterdam 1978).

HENKE, W.L.: Dynamic articulatory model of speech production using computer simulation; doct. diss., Massachusetts Institute of Technology (unpublished, 1966).

KENT, R.D.; CARNEY, P.J.; SEVERED, L.R.: Velar movement and timing: evaluation of a model for binary control. J. Speech Hear. Res. *17:* 470–488 (1974).

KENT, R.D.; MINIFIE, F.D.: Coarticulation in recent speech production models. J. Phonet. *5:* 115–133 (1977).

KOZHEVNIKOV, V.; CHISTOVICH, L.: Speech: articulation and perception (English translation: Joint Publication Research Service, Washington, D.C. 1965).

LINDBLOM, B.: Spectrographic study of vowel reduction. J. acoust. Soc. Am. *35:* 1773–1781 (1963).

LISKER, L.: On time and timing in speech; in SEBEOK Current trends in linguistics, vol. XII (Mouton, The Hague 1974).

MACNEILAGE, P.F.; DECLERK, J.: On the motor control of coarticulation in CVC monosyllables. J. acoust. Soc. Am. *45:* 1217–1233 (1969).

MOLL, K.L.; DANILOFF, R.G.: Investigation of the timing of velar movements during speech. J. acoust. Soc. Am. *50:* 678–684 (1971).

ÖHMAN, S.E.G.: Coarticulation in VCV utterances: spectrographic measurements. J. acoust. Soc. Am. *39:* 151–168 (1966).

PERKELL, J.S.: Physiology of speech production: results and implications of a quantitative cineradiographic study (MIT Press, Cambridge, Mass. 1969).

RAPHAEL, L.J.: The physiological control of durational differences between vowels preceding voiced and voiceless consonants in English. J. Phonet. *3:* 25–33 (1975).

TULLER, B.; HARRIS, K.S.: Temporal models of interarticulator programming (Abstract). J. acoust. Soc. Am. *67:* 564 (1980).

USHIJIMA, T.; HIROSE, H.: Electromyographic study of the velum during speech. J. Phonet. *2:* 315–326 (1974).

Received: September 30, 1980; accepted: October 1, 1980

F. BELL-BERTI, Haskins Laboratories, 270 Crown Street, *New Haven, CT 06510* (USA)